Robust Information Retrieval

Yu-An Liu
Ruqing Zhang
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
{liuyuan21b,zhangruqing}@ict.ac.cn

Jiafeng Guo
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
guojiafeng@ict.ac.cn

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

Beyond effectiveness, the robustness of an information retrieval (IR) system is increasingly attracting attention. When deployed, a critical technology such as IR should not only deliver strong performance on average but also have the ability to handle a variety of exceptional situations. In recent years, research into the robustness of IR has seen significant growth, with numerous researchers offering extensive analyses and proposing myriad strategies to address robustness challenges. In this tutorial, we first provide background information covering the basics and a taxonomy of robustness in IR. Then, we examine adversarial robustness and out-of-distribution (OOD) robustness within IR-specific contexts, extensively reviewing recent progress in methods to enhance robustness. The tutorial concludes with a discussion on the robustness of IR in the context of large language models (LLMs), highlighting ongoing challenges and promising directions for future research. This tutorial aims to generate broader attention to robustness issues in IR, facilitate an understanding of the relevant literature, and lower the barrier to entry for interested researchers and practitioners.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Robustness in IR models, Adversarial robustness, OOD robustness

ACM Reference Format:
Yu-An Liu, Ruqing Zhang, Jiafeng Guo, and Maarten de Rijke. 2024. Robust Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626772.3661380

1 MOTIVATION

Information retrieval (IR) systems are an important way for people to access information. In recent years, with the development of deep learning, deep neural networks have begun to be applied in IR systems [8, 18, 24], achieving remarkable effectiveness. However, beyond their effectiveness, these neural IR models also inherit the inherent *robustness flaws* of neural networks [38, 39, 41]. This poses a hindrance to their widespread application in the real world.

In the past few years, the issue of the robustness of IR has received wide attention, e.g., Wu et al. [42] analyzed the robustness of neural ranking models (NRMs), and a perspective paper on competitive search [21] discussed adversarial environments in search engines. Since then, there has been a lot of work that focuses on different robustness aspects in IR, such as adversarial robustness [23, 25, 27, 41], out-of-distribution (OOD) robustness [12, 38], performance variance [42], robustness under long-tailed data [16], and on the corresponding improvement options. Today, the research community can effectively scrutinize IR models leading to more robust and reliable IR systems.

To ensure the quality of the tutorial, we will focus on the two most widely studied types of robustness issues, namely *adversarial robustness* and *OOD robustness*. There are many analyses and suggestions for improvement around these two robustness issues, but it has not yet been systematically organized. Through this tutorial, we aim to summarize and review the progress of robust IR to attract attention and promote widespread development in this field.

2 OBJECTIVES

1. Introduction. We start by reminding our audience of the required background and introducing the motivation and scope of the robustness issue in IR in our tutorial.

2. Preliminaries. In IR, robustness signifies an IR system's consistent performance and resilience against diverse unexpected situations. There is a large volume of work that covers many aspects of IR robustness, e.g., (i) *Adversarial robustness* [25, 41], which focuses on the ability of the IR model to defend against malicious adversarial attacks aimed at manipulating rankings; (ii) *OOD robustness* [38, 42], which measures the performance of an IR model on unseen queries and documents from different distributions of the training dataset; (iii) *Performance variance* [42], which emphasizes the worst-case performance across different individual queries under the independent and identically distributed (IID) data; and (iv) *Robustness under long-tailed data* [16], which refers to the capacity to effectively handle and retrieve relevant information from less common, infrequently occurring queries or documents.

In this tutorial, we focus on adversarial robustness and OOD robustness, which have received the most attention. Interest in adversarial robustness stems largely from the widespread practice of search engine optimization (SEO) [4]. Concerns about OOD robustness are primarily due to the need for adaptation across diverse

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3661380

and complex real-world scenarios. Moreover, as large language models (LLMs) are being integrated into IR, new robustness challenges emerge; LLMs also offer novel opportunities for enhancing the robustness of IR systems.

Building on these preliminaries, we will cover adversarial robustness, OOD robustness, and robust IR in the age of LLMs.

3. Adversarial robustness. The web is a competitive search environment, which can lead to the emergence of SEO, in turn causing a decline in the content quality of search engines [4, 21]. With the gradual rise of SEO, traditional web spamming [19] started to become an effective way to attack IR systems. However, this approach based on keyword stacking is easily detected by statistical-based spamming detection methods [48].

Adversarial attacks. In order to exploit the vulnerability of neural IR models, many research works have simulated real black-hat SEO scenarios and proposed a lot of adversarial attack methods. (i) First, we introduce the differences between attacks in IR and CV/NLP, including task scenarios and attack targets; (ii) Then, we present adversarial retrieval attacks [1, 22, 25, 29, 47] against the first-stage retrieval models, including the task definition and evaluation. Current retrieval attack methods mainly include corpus poison attacks [22, 25, 47], backdoor attacks [29], and encoding attacks [1]; and (iii) Finally, we introduce adversarial ranking attacks [11, 23, 26, 27, 36, 39, 41] against NRMs with task definitions and evaluation setups. These include word substitution attacks [26, 39, 41], trigger attacks [23, 26, 36], and prompt attacks [11, 33].

Adversarial defense. To cope with adversarial attacks, research has proposed a series of adversarial defense methods to enhance the robustness of IR models. (i) We introduce the objective and evaluation of IR defense tasks. Based on these defense principles, adversarial defense methods in IR can be classified as attack detection, empirical defense, and certified robustness; (ii) We turn to attack detection, which includes perplexity-based, linguistic-based, and learning-based detection [10]; (iii) We present empirical defenses, which encompass data augmentation [9], traditional adversarial training [30, 32], and theory-guided adversarial training [28]; and (iv) We introduce the certified robustness method in IR [40].

4. Out-of-distribution robustness. In real-world scenarios, search engines are in an ever-changing data environment, and new data are often not IID with the training data. Therefore, the ability to generalize to OOD data or not is the basis for the evaluation of IR systems in terms of OOD robustness [42].

OOD generalizability on unseen documents. In IR, the OOD robustness scenarios that have been examined can be categorized into unseen documents and unseen queries. The unseen documents scenario may be caused by adaptation to new corpus [38] or by incrementation of original corpus [2]. (i) Adaptation to new corpus usually refers to the phenomenon that the corpus on which an IR model is trained is not in the same domain as the corpus on which it is tested. Due to the overhead of retraining, the performance of the model on the new domain needs to be guaranteed under zero/few-shot scenario, which is usually solved by domain adaptation [3, 13, 43, 45]; and (ii) Incrementation of original corpus refers to the real-world scenario where new documents are continuously added to the corpus with potential distribution drift. In this situation, the IR model should effectively adapt to the evolving distribution with the unlabeled new-coming data, which is usually solved by continual learning [2, 6].

OOD generalizability on unseen queries. Unseen queries concern query variations [49] and unseen query types [42]. (i) The query variations are usually different expressions of the same information need [34, 49] which may impact the effectiveness of IR models. Many noise-resistant approaches [12, 34, 35, 49, 50] have been proposed for neural IR models; and (ii) Unseen query types refer to the unfamiliar query type with new query intents [42]. Domain regularization [13] is effective for dealing with new query types.

5. Robust IR in the Age of LLMs. (i) We first discuss the potential robustness challenges with applications of LLMs in IR, such as retrieval augmentation [15, 20, 31], and LLMs for ranking [37, 46]; and (ii) Then, we will discuss how LLMs can be used to enhance the robustness of IR systems. These explorations will inspire many novel attempts in this area.

6. Conclusions and future directions. We conclude our tutorial by discussing several important questions and future directions, including (i) There is a diverse focus on the robustness of IR models from multiple perspectives. Establishing a unified benchmark of analysis to systematically analyze the robustness of all aspects of existing models. (ii) For adversarial robustness, existing work on adversarial attacks focuses on specific stages (first-stage retrieval or re-ranking) [25, 41] in IR systems. Customizing adversarial examples to make them effective for all stages is challenging. Therefore, one potential future direction is to explore how we can design a general unified attack method that can cater to every IR stage. (iii) For OOD robustness, the main limitation of existing work is the difficulty of seeing enough diverse domain data in advance, leading to insufficient transfer capabilities of the model. Using the generation capabilities of LLMs to synthesize corpora for adaptation domains seems to be a promising direction.

3 RELEVANCE TO THE IR COMMUNITY

In recent years, a considerable number of tutorials focusing on the topic of robustness have emerged across disciplines within computer science. In KDD'21 [14], CVPR'21 [7], and AAAI'22, there were tutorials on robustness for AI and computer vision. In EMNLP'21 [5] and EMNLP'23 [44], there were tutorials on robustness and security challenges in NLP. The focus of these tutorials was not on search tasks and models.

Search and ranking is a core theme at SIGIR. Evaluation, another core theme at SIGIR, encompasses multiple critical criteria beyond effectiveness for evaluating an IR system. Robust information retrieval aligns well with these core themes. Recently, robust information retrieval has gained considerable attention as more and more work is now devoted to analyzing and improving the robustness of information retrieval systems [17, 23, 38, 41, 42]. Our tutorial will describe recent advances in robust information retrieval and shed light on future research directions. It would benefit the IR community and help to encourage further research into robust IR.

4 FORMAT AND DETAILED SCHEDULE

A detailed schedule for our proposed half-day tutorial (three hours plus breaks), which is aimed at delivering a high-quality presentation within the selected time frame, is as follows:

1. **Introduction** (15 minutes)
   - Introduction to robust IR: motivation and scope
   - Tutorial overview
2. **Preliminaries** (20 minutes)
   - Definition of robustness in IR
   - Taxonomy of robustness in IR
3. **Adversarial Robustness** (50 minutes)
   - Traditional Web spamming
   - Adversarial attacks
     - Comparison: IR attacks vs. CV/NLP attacks
     - Retrieval attacks: definition, evaluation, method, etc.
     - Ranking attacks: definition, evaluation, method, etc.
   - Adversarial defense
     - IR defense tasks: objective & evaluation
     - Empirical defense: adversarial training, detection, etc.
     - Theoretical defense: certified defense, etc.
4. **Out-of-distribution Robustness** (45 minutes)
   - OOD generalizability in IR
   - OOD generalizability on unforeseen corpus
     - Definition & evaluation
     - Adaptation to new corpus
     - Incrementation of original corpus
   - OOD generalizability on unforeseen queries
     - Definition & evaluation
     - Query variation
     - Unseen query type
5. **Robust IR in the Age of LLMs** (20 minutes)
   - New challenges to IR robustness from LLMs
   - New solutions for IR robustness via LLMs
6. **Challenges and Future Directions** (20 minutes)
7. **QA Session** (10 minutes)

5 TUTORIAL MATERIALS

We plan to make all teaching materials available online for attendees, including: (i) Slides: The slides will be made publicly available. (ii) Annotated bibliography: This compilation will contain references listing all works discussed in the tutorial, serving as a valuable resource for further study. (iii) Reading list: We will provide a reading list with a compendium of existing work, open-source code libraries, and datasets relevant to the work discussed in the tutorial. We intend to ensure that all instructional materials are available online.[1] Moreover, we grant permission to include slides and video recordings in the ACM anthology.

ACKNOWLEDGMENTS

This work was funded by the National Key Research and Development Program of China under Grants No. 2023YFA1011602, the Strategic Priority Research Program of the CAS under Grants No. XDB0680102, the project under Grants No. JCKY2022130C039, and the Lenovo-CAS Joint Lab Youth Scientist Project. This work was also (partially) funded by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO), project ROBUST with project number KICH3.LTP.-20.006, which is (partly) financed by the Dutch Research Council (NWO), DPG Media, RTL, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023, and the FINDHR (Fairness and Intersectional Non-Discrimination in Human Recommendation) project that received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101070212.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

[1] https://robust-information-retrieval.github.io

REFERENCES

[1] Nicholas Boucher, Luca Pajola, Ilia Shumailov, Ross Anderson, and Mauro Conti. 2023. Boosting Big Brother: Attacking Search Engines with Encodings. *arXiv preprint arXiv:2304.14031* (2023).
[2] Yinqiong Cai, Keping Bi, Yixing Fan, Jiafeng Guo, Wei Chen, and Xueqi Cheng. 2023. L2R: Lifelong Learning for First-stage Retrieval with Backward-Compatible Representations. In *CIKM*. 183–192.
[3] ZeFeng Cai, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, Xin Alex Lin, Liang He, and Daxin Jiang. 2022. HypeR: Multitask Hyper-Prompted Training Enables Large-Scale Retrieval Generalization. In *ICLR*.
[4] Carlos Castillo and Brian D. Davison. 2011. Adversarial Web Search. *Foundations and Trends in Information Retrieval* 4, 5 (2011), 377–486.
[5] Kai-Wei Chang, He He, Robin Jia, and Sameer Singh. 2021. Robustness and Adversarial Examples in Natural Language Processing. In *EMNLP*. 22–26.
[6] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. Continual Learning for Generative Retrieval over Dynamic Corpora. In *CIKM*. 306–315.
[7] Pin-Yu Chen and Sayak Paul. 2021. Practical Adversarial Robustness in Deep Learning: Problems and Solutions. In *CVPR*.
[8] Ruey-Cheng Chen, Luke Gallagher, Roi Blanco, and J. Shane Culpepper. 2017. Efficient Cost-aware Cascade Ranking in Multi-stage Retrieval. In *SIGIR*. 445–454.
[9] Xuanang Chen, Ben He, Kai Hui, Le Sun, and Yingfei Sun. 2023. Dealing with Textual Noise for Robust and Effective BERT Re-ranking. *IPM* 60, 1 (2023), 103135.
[10] Xuanang Chen, Ben He, Le Sun, and Yingfei Sun. 2023. Defense of Adversarial Ranking Attack in Text Retrieval: Benchmark and Baseline via Detection. *arXiv preprint arXiv:2307.16816* (2023).
[11] Xuanang Chen, Ben He, Zheng Ye, Le Sun, and Yingfei Sun. 2023. Towards Imperceptible Document Manipulations against Neural Ranking Models. In *ACL*. 6648–6664.
[12] Xuanang Chen, Jian Luo, Ben He, Le Sun, and Yingfei Sun. 2022. Towards Robust Dense Retrieval via Local Ranking Alignment. In *IJCAI*. 1980–1986.
[13] Daniel Cohen, Bhaskar Mitra, Katja Hofmann, and W Bruce Croft. 2018. Cross Domain Regularization for Neural Ranking Models Using Adversarial Learning. In *SIGIR*. 1025–1028.
[14] Anupam Datta, Matt Fredrikson, Klas Leino, Kaiji Lu, Shayak Sen, and Zifan Wang. 2021. Machine Learning Explainability and Robustness: Connected at the Hip. In *SIGKDD*. 4035–4036.
[15] Run-Ze Fan, Yixing Fan, Jiangui Chen, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. 2024. RIGHT: Retrieval-Augmented Generation for Mainstream Hashtag Recommendation. In *ECIR*. Springer, 39–55.
[16] Darío Garigliotti, Dyaa Albakour, Miguel Martinez, and Krisztian Balog. 2019. Unsupervised Context Retrieval for Long-tail Entities. In *ICTIR*. 225–228.
[17] Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2020. Ranking-incentivized Quality Preserving Content Modification. In *SIGIR*. 259–268.
[18] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *CIKM*. 55–64.
[19] Zoltan Gyongyi and Hector Garcia-Molina. 2005. Web Spam Taxonomy. In *AIRWeb*.
[20] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research* 24, 251 (2023), 1–43.
[21] Oren Kurland and Moshe Tennenholtz. 2022. Competitive Search. In *SIGIR*.
[22] Zilong Lin, Zhengyi Li, Xiaojing Liao, XiaoFeng Wang, and Xiaozhong Liu. 2023. MAWSEO: Adversarial Wiki Search Poisoning for Illicit Online Promotion. *arXiv preprint arXiv:2304.11300* (2023).
[23] Jiawei Liu, Yangyang Kang, Di Tang, Kaisong Song, Changlong Sun, Xiaofeng Wang, Wei Lu, and Xiaozhong Liu. 2022. Order-Disorder: Imitation Adversarial

Attacks for Black-box Neural Ranking Models. In *CCS*. 2025–2039.
[24] Shichen Liu, Fei Xiao, Wenwu Ou, and Luo Si. 2017. Cascade Ranking for Operational E-commerce Search. In *SIGKDD*. 1557–1565.
[25] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. Black-Box Adversarial Attacks against Dense Retrieval Models: A Multi-View Contrastive Learning Method. In *CIKM*. 1647–1656.
[26] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. Topic-Oriented Adversarial Attacks against Black-Box Neural Ranking Models. In *SIGIR*. 1700–1709.
[27] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Multi-granular Adversarial Attacks against Black-box Neural Ranking Models. In *SIGIR*.
[28] Yu-An Liu, Ruqing Zhang, Mingkun Zhang, Wei Chen, Maarten de Rijke, Jiafeng Guo, and Xueqi Cheng. 2024. Perturbation-Invariant Adversarial Training for Neural Ranking Models: Improving the Effectiveness-Robustness Trade-Off. In *AAAI*, Vol. 38.
[29] Quanyu Long, Yue Deng, LeiLei Gan, Wenya Wang, and Sinno Jialin Pan. 2024. Backdoor Attacks on Dense Passage Retrievers for Disseminating Misinformation. *arXiv preprint arXiv:2402.13532* (2024).
[30] Simon Lupart and Stéphane Clinchant. 2023. A Study on FGSM Adversarial Training for Neural Retrieval. In *ECIR*. Springer, 484–492.
[31] Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. When Do LLMs Need Retrieval Augmentation? Mitigating LLMs' Overconfidence Helps Retrieval Augmentation. *arXiv preprint arXiv:2402.11457* (2024).
[32] Dae Hoon Park and Yi Chang. 2019. Adversarial Sampling and Training for Semi-supervised Information Retrieval. In *The World Wide Web Conference*. 1443–1453.
[33] Andrew Parry, Maik Fröbe, Sean MacAvaney, Martin Potthast, and Matthias Hagen. 2024. Analyzing Adversarial Attacks on Sequence-to-Sequence Relevance Models. In *ECIR*. Springer, 286–302.
[34] Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators. In *ECIR*.
[35] Georgios Sidiropoulos and Evangelos Kanoulas. 2022. Analysing the Robustness of Dual Encoders for Dense Retrieval Against Misspellings. In *SIGIR*. 2132–2136.
[36] Congzheng Song, Alexander M Rush, and Vitaly Shmatikov. 2020. Adversarial Semantic Collisions. In *EMNLP*. 4198–4210.
[37] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *EMNLP*. 14918–14937.
[38] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *NIPS*.
[39] Yumeng Wang, Lijun Lyu, and Avishek Anand. 2022. BERT Rankers are Brittle: A Study using Adversarial Document Perturbations. In *ICTIR*.
[40] Chen Wu, Ruqing Zhang, Jiafeng Guo, Wei Chen, Yixing Fan, Maarten de Rijke, and Xueqi Cheng. 2022. Certified Robustness to Word Substitution Ranking Attack for Neural Ranking Models. In *CIKM*. 2128–2137.
[41] Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. PRADA: Practical Black-Box Adversarial Attacks against Neural Ranking Models. *TOIS* 41, 4 (2023), Article 89.
[42] Chen Wu, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. Are Neural Ranking Models Robust? *TOIS* 41, 2 (2022), 1–36.
[43] Ruicheng Xian, Honglei Zhuang, Zhen Qin, Hamed Zamani, Jing Lu, Ji Ma, Kai Hui, Han Zhao, Xuanhui Wang, and Michael Bendersky. 2023. Learning List-Level Domain-Invariant Representations for Ranking. *NIPS* 36 (2023).
[44] Qiongkai Xu and Xuanli He. 2023. Security Challenges in Natural Language Processing Models. In *EMNLP*. 7–12.
[45] Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. COCO-DR: Combating Distribution Shifts in Zero-Shot Dense Retrieval with Contrastive and Distributionally Robust Learning. *arXiv preprint arXiv:2210.15212* (2022).
[46] Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Are Large Language Models Good at Utility Judgments?. In *SIGIR*.
[47] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning Retrieval Corpora by Injecting Adversarial Passages. In *EMNLP*.
[48] Bin Zhou and Jian Pei. 2009. OSD: An Online Web Spam Detection System. In *SIGKDD*, Vol. 9.
[49] Shengyao Zhuang and Guido Zuccon. 2021. Dealing with Typos for BERT-based Passage Retrieval and Ranking. In *EMNLP*. 2836–2842.
[50] Shengyao Zhuang and Guido Zuccon. 2022. CharacterBERT and Self-teaching for Improving the Robustness of Dense Retrievers on Queries with Typos. In *SIGIR*.